DOCUMENT RESUME

ED 223 715 TM 820 862

TITLE	Test Interpretation, Misinterpretation, and Instructional Planning.
INSTITUTION	Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	82
NOTE	8p.
PUB TYPE	Collected Works - Serials (022) -- Guides - Non-Classroom Use (055)
JOURNAL CIT	SWRL Instructional Improvement Digest; n6 1982
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Educational Planning; Elementary Secondary Education; *Instructional Development; Scores; Teachers; *Test Interpretation; Test Items; Test Results

ABSTRACT

The "Instructional Improvement Digest" communicates advisory information about practical courses of action that can be implemented by teachers and administrators to improve key areas of school instruction. The series digest topics draws upon inquiry associated with the Southwest Regional Laboratory for Educational Research and Development's Proficiency Verification Systems and Services and other pertinent research. The digest seeks to focus on matters of high priority in the conduct of current activities for instructional improvement. This article addresses the matter of how to use student test results for instructional purposes. It provides a simple and practical strategy for using test information sensibly. Interpreting test results, subtest labels, and test items are discussed. Good test consumerism requires judging test items according to the intention of instruction and interpreting scores according to their usefulness in instructional planning. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. S. Escoe

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SWRL
INSTRUCTIONAL IMPROVEMENT DIGEST

No. 6 1982

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED223715

TEST INTERPRETATION, MISINTERPRETATION, AND INSTRUCTIONAL PLANNING

TM 820862

Teachers are required to give achievement tests to students for many different purposes. The intention is always to help teachers and students. Whether teachers and students regard the effort as helpful, however, varies—some do and some don't. One of the ways that everyone *would like* the results to be helpful is in teachers' instructional planning. However, the relationship between test results and instruction seems to be elusive. Is there a secret or mystique between the two?

This article addresses the matter of how to use student test results for instructional planning purposes. But it does not imply that instructional planning can be or should be reduced to a mechanical routine. Such planning inherently must rely on the professional knowledge of the person involved: the teacher. What it does provide is a simple and practical strategy for using test information sensibly.

Seeing "i to i"

Most people would agree that educators should know (1) what they are supposed to teach (intention), (2) what materials and strategies they are going to use (instruction), and (3) how they are going to identify student accomplishments (information). Seeing to it that these three components —intention, instruction, and information— work together is a desirable goal in planning and following curriculum. Though this plan looks good on paper, in practice many things can and do go awry in trying to make it work.

Efforts toward instructional improvement typically begin by assuming that skills, materials, and assessment are coordinated. School staffs put a lot of time and energy into devising and carrying out the improvement plan, and they expect to see improvement in students' test scores. But the results may not reflect the effort. Sometimes this is because the improvement effort itself was conceived hastily. Other times it is because staff expectations were unreasonably high. Most often, however, a post-mortem reveals that the problem was a lack of coordination between the underlying components—intention, instruction, and information. A case study may help to illustrate the situation.

A Case Study

All the fifth-grade teachers in a school met to review their pupils' scores on the district's competency test. They wanted to use the information from the test to plan improvement for their instructional program. Looking at results from the Composition part of the test, they noted that their students' performance in "Mechanics" was relatively low. They decided as a group to make Mechanics a priority area for improvement in the coming school year.

So the leader of the group wrote the word "Mechanics" on the chalkboard and asked what skills should be included. The responses from the other teachers were global: capitalization, punctuation, paragraph indentation, spelling, etc. At this point, the teachers could have proceeded in one of two ways: they could have decided to try to improve instruction relative to their list, or they could have paused to check their list against the district's curriculum guide for Grade 5 Composition instruction. Had they taken the second path, they would have found the following in the district's guide:

1. *Capitalizes the first letter in titles: Mrs., Miss, Ms., Mr., and Dr.*

2. *Uses periods at the end of abbreviations and initials.*

3. *Capitalizes the first, last, and important words in a title.*

4. *Uses commas in quotations.*

5. *Uses commas to separate items in a series.*

6. *Uses commas between city and state.*

If the teachers had taken the first route, that is, defined their intentions on the basis of the more global list, they could have been disappointed in their test scores for Mechanics at the end of the year since the instruction provided might not have matched the more specific skills assessed. If, on the other hand, the teachers *had* followed the more clearly marked instructional path, or at least had assured that these skills were included in instruction, they would be more apt to see improvement in the test results.

The point of the illustration is that it is very important to understand the *specific* instructional expectations in order to plan for effective improvement. Not all districts and schools list their expectations in such detailed form as the example. In such cases the intentions and the instructional plans may both have to be inferred from the assessment information. However, there are pitfalls in trying to infer such meaning from tests.

Pitfalls in Interpreting Test Results

Teachers face two major pitfalls in using test results for instructional planning: (1) interpreting subtest labels and scores and (2) interpreting individual test items. A Composition test, for example, may consist of subtests such as Sentence Processing, Paragraph Development, Mechanics, etc. Can we tell from these labels exactly which skills are assessed under each heading, wherever the heading is used? For instance, is "using commas" included under Mechanics in the third-grade test? Is it included under the same heading in the fourth-grade test? Does it appear at all in the fifth-grade test? By interpreting performance on individual items, we can find out how the items were answered by one student, by a class, a grade-level, a school, and even an entire school district. These statistics are easy to get, but what do they tell us about how students write? Let's take a closer look at both the labels and the items.

Interpreting Subtest Labels

Achievement tests for elementary school students often are organized by grade level. For example, there is a Grade 1 Mathematics test, a Grade 2 Mathematics test, etc. The same holds for other subject areas such as Reading, Composition, Science, etc. Commonly, each subject area test is composed of several subtests, for example:

The Instructional Improvement Digest communicates advisory information about practical courses of action that can be implemented by teachers and administrators to improve key areas of school instruction. The series of Digest topics draws upon inquiry associated with SWRL's Proficiency Verification Systems and Services and other pertinent research. The Digest seeks to focus on matters of high priority and concern in the conduct of current activities for instructional improvement. Suggestions of topics for future inclusion in the Digest series are invited, and may be directed to Adrienne S. Escoe, Editor, Instructional Improvement Digest, Southwest Regional Laboratory for Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, California 90720.

Preparation of Instructional Improvement Digest is part of SWRL's inquiry into Schooling Practices and Effects, which is supported by a grant from the National Institute of Education, Department of Education. However, the opinions expressed here do not necessarily reflect the position or opinion of NIE, and no official endorsement by NIE should be inferred. **(The Digest may be duplicated and distributed to interested educators.)**

Mathematics

- Number Recognition
- Computation
- Measurement
- Problem Solving

Reading

- Decoding
- Vocabulary
- Sentence Comprehension
- Paragraph Comprehension

Composition

- Sentence Processing
- Paragraph Development
- Mechanics
- Spelling

Though the headings are usually the same for tests and subtests at each grade level, e.g., "Mathematics" and "Measurement," the skills assessed may be very different. But relying on headings alone can be misleading in interpreting test results. *The pitfall is overlooking the DIFFERENCES and RELATIONSHIPS between the same labels at different grade levels.* Here are some situations that illustrate the pitfall:

Situation 1: Same label but different, unrelated meanings

Grade 3 Measurement items assess recognition of the value of different money denominations— e.g., penny, nickel, dime, quarter, half-dollar, and dollar.

Grade 4 Measurement items assess knowledge of metric and nonmetric units—e.g., measuring length to the nearest centimeter, meter, inch, and foot.

Clearly, the Grade 3 and Grade 4 measurement skills are different and fairly unrelated. Moreover, these Grade 3 measurement skills are probably not prerequisite to the Grade 4 measurement skills. In other words, a student probably doesn't have to have the Grade 3 measurement skills (i.e., money) in order to be successful in learning the Grade 4 measurement skills (metric/nonmetric). For instructional planning, this situation implies that the illustrated Grade 3 measurement skills do not have to be in place prior to teaching the Grade 4 measurement skills. Where monetary and metric measures are taught is discretionary. But unless the instruction matches the assessment and vice versa, little improvement is likely to follow from the instructional plan.

Situation 2: Same label but semi-related meanings

Grade 4 Mechanics items assess the use of apostrophes in singular possessive forms. For example:

> Jenny has an uncle. Jenny likes to visit her _____ house.
>
> a. uncle's
> b. uncles
> c. uncl'es

Grade 5 Mechanics items assess the use of commas to separate items in a series, for example:

> There were lions__ tigers, and elephants at the circus.
>
> a. !
> b. ,
> c. .
> d. none of these marks

Clearly, the Grade 4 and Grade 5 Mechanics skills are different and neither is prerequisite to the other. However, both are probably required for the student to write a satisfactory story or composition in the fifth grade. For planning instruction, this situation implies that both skills probably need to be in place for a student to write a Grade 5 composition; however, the Grade 4 skill (apostrophes) does not necessarily have to be in place prior to teaching the grade 5 skill (commas). Rather, either one of these skills can be taught first or the two may be taught concurrently.

Situation 3: Same label with direct, prerequisite meanings

Grade 3 Multiplication items assess multiplication facts through 9.

Grade 4 Multiplication items assess the multiplication algorithm involving up to a three-digit multiplicand and up to a two-digit multiplier.

The Grade 3 and Grade 4 skill area labels are the same—i.e., "multiplication." The Grade 3 multiplication skill (multiplication facts) is a direct prerequisite to the Grade 4 multiplication skill (multiplication algorithm). A student should do well on the Grade 3 multiplication skills in order to be successful in learning the Grade 4 multiplication skills. This situation implies that for students who are not skilled in Grade 3 multiplication, the teacher should plan additional instruction before teaching the multiplication skills designated for Grade 4.

In summary, each subject area is represented by a subtest reflecting skills across grade levels that may be related in three different ways, each of which can present a pitfall to teachers. While we often presume that most situations are of the third type—i.e., the skills assessed at one grade level are direct prerequisites to the skills of the next grade level, that is not always so.

One reason that the other two situations are overlooked is that many of the record-keeping devices or charts tend to hide the relationships. For example, charts on which test performance is recorded by hand are usually simple grids. The system for using the grids is basically very easy to follow. Teachers mark the box with a slash (/) when the student is working on the skill, and cross the slash into an X when the student has "mastered" the skill. Usually skills are taught in the same left-to-right order as listed in the chart. Two features of this system tend to conceal the relationships between skill areas and subtests:

Student Name	Skills			
	Skill 1	Skill 2	Skill 3	...

1. All the boxes in the typical grid are of the same size even though the listed skills are of different "sizes." For example, one box may represent the Grade 3 skill "recognizes pennies, nickels, dimes, and quarters." Another box of the same size on the Grade 3 grid may represent the skill "solves word problems involving two-digit addition or subtraction."

2. A left-to-right check-off sequence on the grid tends to hide the interrelationships between skill areas and subtests. That is, the left-to-right sequence tends to imply that a left-side box is a prerequisite to the right-side box, which may or may not be true.

Interpreting Test Items

Items are the basic building blocks of tests. The information yielded by a test is only as useful as the information provided by each test item. So let's see what test items can tell us about students' skills.

In the preceding section, we found important differences between subtests across grade levels. In the same way, two test items may have identical labels but still show important differences.

This situation becomes clearer when we compare items that assess instruction indirectly with direct assessments. Two examples illustrate important differences between these item types.

Example 1: Grade 2 Measurement (Telling Time)

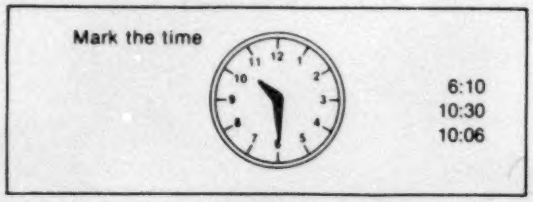
a. **Indirect** instructional item (answered correctly by 60% of Grade 2 students)

Mr. Baker washed his car. The two clocks show you when he started and when he finished. At what time did he finish?

START FINISH

8:40
8:30
8:00
9:30

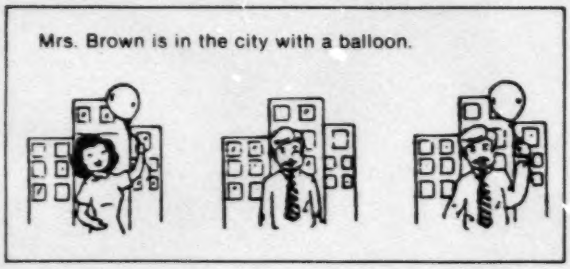
b. **Direct** instructional item (answered correctly by 80% of Grade 2 students)



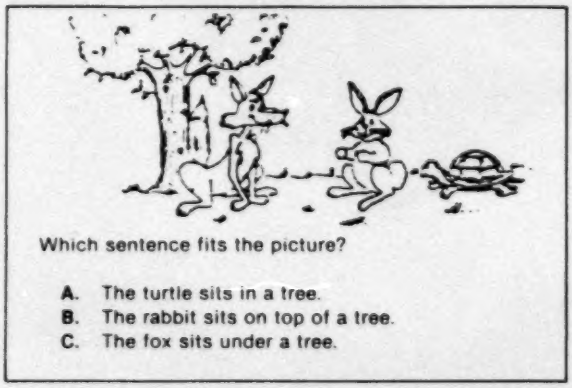
The indirect instructional item contains extraneous material; to answer the question correctly, the student doesn't need to see the START time. The direct item provides better planning information. It eliminates unnecessary distractions and focuses instead on whether students have learned the skill of telling time.

Example 2: Grade 2 Sentence Comprehension

a. **Indirect** instructional item (answered correctly by 50% of Grade 2 students)



b. **Direct** instructional item (answered correctly by 73% of Grade 2 students)



The direct assessment item requires more reading than the indirect item. Though the indirect item is brief, there are two elements that make it more difficult for children. First, the appellation "Mrs." is typically part of oral vocabulary taught in Grade 2 but not a part of the *reading* vocabulary taught in that grade. Second, in practice materials such as workbooks, students are often asked to mark the "one that is different." Thus, some students may automatically mark the picture *without* the balloon.

An incorrect response to the first item may indicate that a student (a) really hasn't learned the target skill (that is, comprehending a written sentence), (b) doesn't understand the item format, or (c) is confused by the illustration, and so on. If students answer the second item incorrectly, at least teachers can be more confident that a student was truly weak in a tested skill and did not respond incorrectly because of the nature of the test. For planning purposes, the direct item provides information that is amenable to instruction. The more indirect the item, the less clear the implications for instructional planning.

The Question is "What is the Question?"

In interpreting test results for planning instruction, there is one basic question that must be consistently asked. It is:

Are these test questions something my students have seen or practiced in their classroom work?

If test information is to be useful for instructional planning, it must be strongly related to actual classroom practice. The examples in the earlier part of this article—telling time and comprehending sentences—illustrate how important this relationship is. These examples show how some versions of a test item provide better information for planning instruction than other versions of the "same" item. Consequently, good test consumerism or test-wiseness requires critical review and examination of the items in a test. Serious consideration of the basic question in the box above constitutes a "critical review." Two variations of this basic question are examined in the remainder of this article.

Variation 1. Are these the same questions I've been asking my students? The "same" question can be asked in different ways. We as adults and teachers may see two questions as the "same" question. However, children may see them as different questions. For example, two typical and roughly equivalent reading comprehension questions are

Question 1: What is this story about?

Question 2: What is the main idea?

Without explicit instruction, third- and fourth-grade students may well understand one question and not the other. That is, they may be able to tell you what the story is about but not be able to tell you the main idea because they are not familiar with the phrase "main idea." Conversely, students may not understand that "What is this story about?" is a request for a *central* theme, not just a detail of a story.

As another example, two mathematically equivalent addition problems that occur frequently in tests are

Question 1: 9 + 4 + 7 = ?

Question 2:

$$\begin{array}{r} 9 \\ 4 \\ +\ 7 \\ \hline ? \end{array}$$

Children perform differently with horizontal and vertical formats in addition problems. Some children don't see a horizontal format except on tests. (In fact, some *people* don't see horizontal addition formats except on tests!) Some children may be able to correctly add 9, 4, and 7 but not realize that it is the same question in the horizontal format. The point, again, is to ask whether the test question looks like the instruction students have been used to seeing.

Variation 2. Does the mix of test items accurately reflect the breadth and depth of my instructional program? Do the *kinds* of skills covered in the test represent the mix of skills taught in my program? Does the *number* of items devoted to each skill category represent the relative importance or amount of time spent in instruction? Just as you expect well-written unit tests or chapter tests to "mirror" the unit or chapter, so should you expect semester, year-long, or multi-year tests to mirror the instruction covered in the respective period of time.

Instructional Planning

Two major steps are involved in "using" test information for instructional planning. Together, these steps summarize most of the implications of the preceding discussion. If the test content coordinates with your instructional program, proceed directly to Step 2. If the test content doesn't, you have three choices.

Step 1. Analyze the test *content* item by item and subtest by subtest to establish its conformity to your instructional program.

a. Modify your instructional program so that it better matches the test.

b. Disregard those portions of the test that do not match your instruction.

c. Work toward coordinating the test and the instructional program (i.e., work toward modifying the test and the instructional program).

Step 2. Look at the *numbers*. That is, look at student performance.

If student performance is good, then

a. Continue as before, or

b. Consider doing less. Students may have already learned the test content through other instruction or through other means (e.g., home or TV).

If student performance is low, then

a. Consider doing more. Spend more time teaching the skill area. It's likely that the content is simply not being "covered" adequately.

b. Consider doing instruction differently. Change teaching strategies or materials.

c. Consider doing less. It may be that the content area is not "worth" the instructional effort, given other instructional needs.

d. Continue as before. It may be that the content is dependent upon some other content that wasn't taught adequately. Thus, attending to teaching the related or prerequisite content may be what is needed.

Two observations are in order regarding the preceding choices. One is that we must exercise care when deciding to "raise" low scores. Raising a score from 80% correct to 85% correct may well take more instructional effort and time than raising a score from 35% to 70% correct. Since a score of 35 basically represents "no knowledge," the job of moving from 35 to 70 represents the job of teaching something to somebody who doesn't know very much about the something to begin with. In other words, this is a fairly typical instructional job.

On the other hand, raising a score from 80 to 85 is an effort in "fine tuning." A score of 80 represents a fair amount of knowledge. Raising the score to 85 may mean removing careless errors from the performance. For example, "teaching" students *to be more careful* in the long division process is fine tuning and is different than teaching them *the process*. Trying to remove the arithmetic errors from the long division process may take more time than teaching the process itself. (Indeed, just getting students to use long division in other applications can provide practice in fine tuning.)

The second observation is that all four choices are really dependent upon knowing the substance and structure of instruction. *That*, after all, is the "secret" of instructional planning. Test labels and scores should support this understanding of instruction; they can be interpreted only relative to the substance and structure of the instructional program. In short, teachers should base their interpretation of test information on what they know best—their instructional program.

Summary

The basic lesson in using test information for instructional planning is that we must get behind the labels in order to intepret tests and test items properly. If tests are used to provide information on the effectiveness of instruction or a school improvement effort, then the items represent a concrete and functional definition of the intention of instruction. It is, therefore, extremely important that the items chosen are coordinated with the intention of instruction.

Going only by general labels such as "Measurement" or "Sentence Comprehension" will not assure the desired results. All tests are not the same, despite the fact that they may carry the same labels. Good test consumerism, i.e., test-wiseness, requires comparison shopping for tests and correct interpretation of test scores. Items must be judged according to the intention of instruction, and scores must be interpreted according to their usefulness in instructional planning. Doing anything less will yield results that don't reflect the professional time, effort, and commitment put into an instructional program. They won't show fully what students, teachers, and districts have accomplished.

—George Behr
Senior Member of the Professional Staff
SWRL Educational Research and Development

Note: Content of this article is drawn from a series of technical reports on student accomplishment information systems, written by Aaron Buchanan, Patricia Milazzo, and Richard Schutz.

Readers' comments are always welcome.